DOCUMENT RESUME

ED 347 200 TM 018 718

AUTHOR Riegel, Laurel Blyth
TITLE Texas Campus Report Cards: The Story of One Blind Man and Six Elephants.
PUB DATE Apr 92
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Accountability; Accreditation (Institutions); *Campuses; Comparative Analysis; Educational Assessment; Elementary Secondary Education; *Evaluation Problems; Excellence in Education; Institutional Characteristics; *Institutional Evaluation; Management Information Systems; *Profiles; State Programs
IDENTIFIERS Educational Indicators; *School Information System; *Texas; Texas Education Agency

ABSTRACT

The Texas Education Agency prepares school report cards (SRCs) as part of its academic excellence indicator system (AEIS) of local school accountability. The SRCs are used for evaluating campuses, determining accreditation status, and identifying and rewarding campuses for outstanding performance. Condensing the complex data into the few statistical indexes that a SRC contains is a formidable task. It is argued that the design and revision of the AEIS requires more energy than it has received to date. The methods the AEIS uses for grouping campuses for comparison are described. Shortcomings and recommendations for a campus-wide accountability system are enumerated. Current suggestions and opinions expressed by educational researchers about assessment and evaluation are reviewed. A summary is provided of selected demographics of the changing population of America and implications for education. A paradigm shift for assessment and evaluation is proposed. The Texas SRCs do not provide the quality or type of accountability information needed for the tough decisions facing education today. Accountability for programs, schools, and teachers must reflect the actual functions and goals in operation for the benefit of each learner. Five tables present study data. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

LAUREL BLYTH RIEGEL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED347200

Texas Campus Report Cards: The Story of One Blind Man and Six Elephants

by

Laurel Blyth Riegel, Ph.D.

Coordinating Director – Research and Testing

Richardson Independent School District

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 20, 1992

TM018718

 BEST COPY AVAILABLE

Texas Campus Report Cards: The Story of One Blind Man and Six Elephants

The latest incarnation of local school accountability in the State of Texas is the *academic excellence indicator system* (AEIS) and the collateral school report cards. Based on a very small set of variables, campuses are grouped across the state for purposes of comparison among themselves on a modest set of selected outcome variables. Clusters of schools are identified and group averages within each cluster are calculated for each outcome indicator. Comparisons with group averages yield individual school *report cards*, available for the 6,700+ campuses statewide. All data utilized in the AEIS and on the report cards are collected and reported by the state department of education.

The intent is to provide an annual description of campus progress towards stated campus-performance objectives. The result is a high-stakes enterprise. The subsistence of schools and their staffs hinges upon the validity of the data set and procedures used. Can the complex data set be reduced to a few indices and still distinguish successful schools?

Senate Bill 1 became law in the State of Texas on June 6, 1990. Prior to its being passed in January of 1990, an advisory committee was organized to formulate a system of variables which would indicate *academic excellence*. This body, the State Board Academic Excellence Indicators Advisory Committee, reviewed and recommended variables to be used to group "like" campuses. They also selected outcome measures for the proposed campus grading system. A second committee, appointed by the Governor, focused specifically on acceptable measures of excellence at the campus level.

The applications for the academic excellence indicator system encompass the following:

- Evaluating campuses annually.
- Determining district accreditation status.
- Identifying exemplary campuses.
- Rewarding campuses financially for outstanding performance.

The Texas Education Agency (TEA) touts the school report cards as the best accountability measures available for elementary and secondary public education in the state. However, to condense the complex data describing both the similarities among campuses and the desired outcomes of public education into such a few statistical indices is a formidable task.

The statistical complexity of the problem stems from seemingly conflicting goals. The model must include as much accurate, valid, available information which describes, impacts, and contributes to a school's persona. Yet, the data set must remain streamlined to enable concise and simplified school profiles to be drawn. The statistical procedures used for grouping and grading outcomes must also be sensitive enough to identify critical indicators. At the same time, the methodologies must not be so obtuse as to defy explanation or interpretation to the general public.

TABLE 1 STEPS REQUIRED TO DESIGN SCHOOL REPORT CARDS
Identify variables to use as grouping indices.
Determine grouping methodology.
Select criterion or outcome indicators.
Design methodology to predict annual, expected performance levels for each campus.
Design presentation format.
Design "grading scale" and establish criterion scores (for assigning letter grades).
Develop "training" materials for statistically nontechnical audiences.

Problem Statement

The creation of a viable school report card is a multivariate problem with seven basic dimensions. Table 1 lists the design steps involved in order of execution. To the extent that inadequate results stem from any one of these steps, the resulting report cards are at best misleading as summaries of campus success or failure.

It is this author's opinion that more energy is needed in the design and revision of the AEIS than has currently occurred, *prior to* its publication. In fact, what may be needed even more is a re-examination of the fundamental focus of campus evaluation and accreditation. Administrators, legislators, and evaluators may all be blindfolded and isolated by convention and tradition. Each of them in solitude grapples with a small herd of elephants, clutching at tails, tusks, and trunks. Each tries to explain, define, contrast, evaluate, and predict from the amalgam of data to a unified description of the beast. It is no wonder that individual assessments often fail to reflect each other's conclusions.

A discussion of the methods and results of the AEIS from the State of Texas initiates this paper. Following this, the shortcomings and recommendations for a campus-accountability system are enumerated. Next come current suggestions and opinions expressed by educational researchers, regarding the role of assessment and evaluation in the school of the twenty-first century. This is succeeded by a summary of selected demographics of the changing population of America and implications for the role of education. The closing section proposes a paradigm shift for assessment and evaluation that parallels the "break-the-mold" model for American educational reform.

The Texas AEIS: Grouping Campuses for Comparison

Originally, the advisory committees selected four variables which correlated highly with proposed outcome measures:

- Percentage of students participating in free- and reduced-lunch program.
- Percentage of limited-English-proficient students.
- Percentage of minority students.
- *District* wealth.

In addition, the legislature required in Senate Bill 1 that mobility be included as a grouping variable, although there are no mobility data available on a statewide basis. More will be said about this shortcoming later. For now, the four variables listed above were weighted 70%, 13%, 12%, and 5%, respectively, in a regression formula to create an index

for each campus. Campuses were then rank-ordered from highest to lowest value on this index.

The comparison group for an individual campus consists of the 50 schools preceding it and the 50, following, in the ranked listing. Thus, each group consists of 100 schools. Obviously, the AEIS argument for comparability among grouped campuses relies exclusively on the variables selected and the adequacy of the statistical procedure used to produce the grouping index. Several concerns exist with the validity of the grouping results. Are the variables nonredundant? Are they conceptually meaningful? Are they practically as well as statistically significant?

Nonredundant. Grouping variables should fan out the population to be compared. The population ideally falls into a number of identifiable clusters. These clusters should lend themselves to meaningful interpretation regarding campus similarities and differences. For example, this author's research has found several available grouping variables to be highly correlated and, therefore, redundant. Socioeconomic status[1], ethnic-minority membership, and limited-English proficiency correlated among themselves from .78 to .94 at grades three though twelve in longitudinal studies. It would seem that any one of these variables would serve as a "spokesperson" for the entire grouping of them. To use all of them is to tell relatively the same story again and again, incorporating little if any new information in each reprise.

Conceptually meaningful. Variables used to group schools must possess a great deal of face validity. If school staffs and the general public see no connection between a variable and school operation, utility is greatly diminished. For example, student turnover throughout the year is accepted as a critical factor in school operation. However, the extent to which it is accepted as a useful grouping factor depends on its definition and its accuracy as collected.

In addition, the grouping design should incorporate at least some factors which are controllable. There are a number of predisposing demographic characteristics that describe school environments. So many of these predisposing factors, however, are not subject to the control of the school staff nor the community. To the extent that the majority of grouping factors are beyond experimental control, the grouping is a descriptive profile with no operational access.

Statistically and practically significant. With extremely large samples, small differences may be statistically significant. We also desire that differences be practically significant, in terms of predicting progress or need for improvement. For example, slight differences of one or two percentage points in dropout rates might qualify as statistically significant, but is this practically significant? Negligible differences should not receive inflated attention simply because large sample sizes lead to statistical significance. A variety of techniques should have been used to investigate the relationships among contending grouping variables. Confirmation across several techniques of statistical significance adds to face validity.

Currently, the list of grouping variables has changed to include a measure of student mobility. The weightings are also different:

[1] Research to date defined campus socioeconomic status by proportion of students receiving free- or reduced-price lunches.

- Percentage of students participating in free- and reduced-lunch program (40%).
- Percentage of minority students (40%).
- *District* wealth (10%).
- Percentage of limited-English-proficient students (5%).
- Percentage of student mobility (5%).

At least three of these five critical factors in grouping campuses for comparisons are redundant: percentages of free- and reduced-lunch participants, of minority students, and of limited-English-proficient students. There is a question of practical validity in vesting 80% of the outcome of the grouping index upon two highly correlated independent variables. One of the other grouping variables, district wealth, is incapable of distinguishing among all campuses within the same district. The mobility measure currently used is derived from *self-report* on the statewide assessment during the fall of the academic year and deserves more attention.

Each fall, students are asked during the statewide assessment to indicate on their answer sheets how long they have been enrolled in their respective districts. The validity of this factor is doubtful, given this author's first-hand experience with the confusing answers from asking the same question of parents upon first-time enrollment in a school. In fact, state agency staff has admitted as much. In addition, schools which do not contain the grade levels tested with the fall assessment device may have *missing data* for this variable. In this case, the average of the other four variables is substituted. For all campuses within six special statutory districts, the district wealth is not available. Similarly, the other four variables' average is substituted in these cases. However, for schools missing both district wealth and mobility for the reasons stated, both variables are replaced by the average of the *remaining three variables*.

Concerns also exist regarding the appropriateness of the definitions selected for variables other than mobility. For example, the proportion of students participating in a federally-funded free-/reduced-lunch program is not the best indicator of socioeconomic status, for a number of reasons. Data such as parent income level, median income per neighborhood or campus attendance area are admittedly evasive. Self-report generally produces suspect if not downright confounding results. It is not obvious that poorly defined available data are better than no data at all, in this case. Could there be a linkage between tax files and TEA's database, given assurances of maintenance of privacy?

Beyond the questions raised by the selection of grouping variables lies the issue of choice of statistical technique. There certainly is no best method for determining clusters of like campuses. There may, however, be a number of worse methods. There are nearly as many techniques for consideration as there are statisticians in the State of Texas. Rather than debate the legitimacy of one method over the other, perhaps it makes more sense to employ several methods and use the results which represent a consensus. Confirmatory analyses, at least, would be helpful.

The Texas AEIS: Comparing Campuses on Critical Outcomes

The criteria selected for grading performance determines the picture of success vs. failure of any enterprise. In high-stakes situations, it is essential that the criteria used represent a realistic

sampling of the domain originally defined as the desired outcomes.

Significant outcomes. Understandably, vast effort and energies are dedicated to increase performance levels at each campus in the state. Budgets hinge upon year to year assessments of success. The selected outcomes should be those which we desire. High school outcomes include certain test scores:

- *Texas Assessment of Academic Skills* (TAAS).
- *Texas Academic Skills Program* (TASP).
- *Scholastic Aptitude Test* (SAT).
- American College Testing (ACT) Assessment.

Also included are attendance and dropout rates, proportions of students receiving advance seals on transcripts and enrolled in advanced courses, and expected graduation rate. Do these outcomes adequately reflect the operational objectives of campuses? Is the statewide assessment device an appropriate and sufficient measure of state curriculum as it is defined locally?

The intention is to collect information *in the future* on other aspects of public school performance. Some of the to-be-collected information may be more useful in predicting a school's success rate with the changing student population of the late twentieth century. More will be said about realistic outcomes in a later section.

Outcomes desired by "clients" (community and students). Similar to the last concern is the question, do the clients of public education (community and students) agree that the outcomes selected relate to their perceptions of success? Does the business community really care how many students graduate with advanced seals on transcripts? Do parents care about the attendance rate of their child's school? Are the scores on statewide assessment devices significant to parents and students, aside from the punitive aspect of failing the exit level? Are the average SAT or ACT scores of the local school truly a critical concern to the typical taxpayer?

There may be some way to salvage these outcomes by referencing them to a smaller subgroup than that of the entire school. The subgroup for reporting needs to be an appropriate one, for example, only college-bound students for the SAT and ACT scores. Even so, more realistic success indicators are needed to address daily life in the late twentieth century for the typical student leaving public education.

Positive outcomes desired by students. The outcomes desired by a student might not be those conjured up by legislators and administrators. Students might wish to be certifiably employable. They might wish to be assured the ability to earn at least a parttime livelihood while continuing their education. They might want certain skills required of modern American citizens which do not easily sort out into subject-area categories. We talk a great deal about lifelong learning now, especially with the aging work force. What are the real goals for lifelong learning? Do the SAT, ACT, or other standardized tests assess one's mastery in this area?

Outcomes related to school goals and objectives. Site-based leadership is rapidly becoming the norm in districts across the nation. This is particularly true in Texas, where this form of school management has been legislated. How do top-down state and district goals weigh up against local campus priorities? Whose wishes are more important: those belonging to the state or to the local community? If state-level goals are predominant, is the edict to move

towards site-based leadership a definite conflict? Or is the statewide thrust towards this form of school management really a mixed signal?

Outcomes related to the avowed intent of education. Education is a twenty-first century survival skill. What are the intended outcomes of the provision of education to students? Are we planning to produce good citizens, and do we have consensus on how that outcome is defined? Are we trying to produce lifelong learning skills, or a renewed ardor for knowledge? Do we want graduates of our system to be savvy consumers in a modern, pluralistic, technological culture? Is the priority on producing wage-earners and taxpayers? Our offspring are ultimately responsible for the future of our society, government, economic status, and general national welfare well into the next century. Is it unreasonable to ensure successful management of that not-so-distant future through the public education of our youthful masses?

So many stated goals of education seem to extend just a few years past high-school graduation. In a later section of this report, current information is cited which indicates that the rapidly changing demographics of America present serious issues regarding how to plan for the needs of the citizenry. It is insufficient to monitor simply graduation status and test performance within the few years after high school. We might want to ensure that the future citizens of America are employable, responsible economically and socially, and capable of managing the privileges and obligations of a democratic union. How else do we ensure these outcomes other than to guarantee their inculcation through the educational system?

Multidimensional picture of campus progress. The school report cards supposedly serve as profiles of campus climate, persona, and progress. Only two outcomes currently included focus upon noncognitive aspects of the educational experience: dropout and attendance rates. Surely this presents a lopsided view of a culture which involves its participants for a third to a half of their time each year, teachers and students alike. What are the desirable noncognitive outcomes? Are they the same all across the state? If some are applicable statewide, should they be standardized in definition and metric? Is this another site-based issue?

Process as well as product measures. The key to evaluating institutions which claim to be change agents is to monitor process as closely as product. Product measures are often indicative of change being needed or having taken place. For example, blood pressure, basal temperature, skin flush, and pupil dilation are all indicators of change in human physiology. None of these alone or in concert identify the underlying cause for change.

Likewise, a campus is an organic system. Monitoring only the surface changes such as test scores, graduation and attendance rates, and enrollment in advanced programs is completely inadequate as a measure of functionality, good or bad. Admittedly, process evaluation requires more resources in terms of time, staff, and money than most product evaluation. It may be the only way to highlight differences which distinguish excellent campuses from those in trouble. Important process variables need to be identified, clearly defined, and then measured for baseline determination.

Longitudinal basis and cohort disaggregation. In order to monitor progress, a longitudinal frame of reference is required. Only on certain outcomes are "current year" and "prior year" noted on the AEIS pages. The AEIS pages themselves are relatively useless for long-term monitoring while they contain only one year's data for most variables, both grouping and outcomes. Campus staff must perform paper comparisons, year after year, or re-enter the AEIS information themselves into a computerized system for analysis. This introduces clerical error and additional paperwork burden at the campus level. At the district level, the errors may be multiplied by at least the number of campuses in the district.

A pilot of the AEIS and the school report cards was never conducted. Legislation required immediate implementation after a very short timeline for planning and design. In most cases, pilots serve as testing grounds for new systems. In *this* case, we have no idea of the baseline for the statistics presented in the AEIS. We also have not studied the appropriateness of the methodologies used or of the effect of the report cards themselves on the public, school staffs, and administration. A well conducted, two-year pilot would have highlighted any discrepancies in the variable selection, methodology, and overall design of the report cards. As it is, the discrepancies and shortcomings are being discovered as the information is presented to the public and interpreted at large.

Evaluation should address performance of students who have been subjected to a school's influence for a minimum period of time, for example, two years. If campus averages include the performance of students who just enrolled, their interpretation is spurious. Many districts in Texas share the problems associated with high mobility. As mentioned earlier, there is no adequate statewide index for student turnover among districts. When averaged across all students enrolled, the outcome measures by campus are attenuated by the effects of student mobility. The impact of large proportions of non-English-speaking, highly mobile students within some districts renders utility of group averages null and void, statewide as well as for individual campuses. This does not mean that highly mobile subpopulations should be ignored in accountability analyses. It does imply that substantial intergroup differences must be accounted for in the analytical design *before* the numbers are analyzed and published.

Validity and stability of outcomes and grouping variables. There is no available information about the experimental interaction among the grouping and outcome measures. In fact, TEA states that ongoing analysis and extension of both variable sets are planned. Rather than introduce flexibility into the model, this ongoing evolution will likely provide obstacles to year-to-year comparisons of the information. For example, if TEA sculpts the mobility statistic across the next several years in order to produce a "valid" index, the group outcome averages will hardly be comparable from one year to the next.

The real question of validity of the outcomes is also of concern. As a nation, we place far too much emphasis on test scores because they are available. We in education in the State of Texas must clarify our vision. Based upon this vision, we must formulate the goals and objectives of educating the public. Is our true goal in Texas to have high SAT scores? Do only successful campuses have an average daily attendance of 96% or

higher? Is there something else more crucial to successful functioning in current society besides these outcomes? Would the resources be better allocated to defining those alternatives and designing measurement strategies for them? Otherwise, we merely evolve the wheel year after year, hoping that it finally approaches something that rolls more or less steadily.

The Texas AEIS: Interpretation of Information

One of the arguments for the simplicity of the variable set, statistical methodologies, and format is that the information will be easily interpreted by nontechnical audiences. This assumption bears debate.

No guidelines exist for the usage of the school report cards in prescribing needed changes for unsuccessful campuses. Teachers glean no specific information from the report other than their school as a whole scores low or high. Even the definitions for "low" and "high" are unclear. What magnitude of difference between the campus average and the group average is significant?

Significant differences would need to be calibrated for all outcomes, and they might drift from year to year. They probably are not equal throughout the ranked listing of campuses. Lower and higher portions of the ranked listing would likely have different standard errors than the middle portions. What about changes for a specific campus from one year to the next? What is significant?

All of this is aside from the question of practicality. How do the campus staff and faculty interpret being "above average" in their grouping, despite reflecting extremely low outcome levels? How do they explain to their parents and students that the campus received an "A" from TEA even though the clientele of the school would give it an "F"? If we truly strive for excellence and believe that *everyone* can learn, so what if a campus average is lower or higher (no matter *how* much) than that of the average of another hundred campuses? If the group comparisons really are important, how can the ethnic differences within a school be interpreted when no group ethnic profiles are provided?

Progress from one year to the next for individual campuses is much more easily understood and trackable. Parents, teachers, administrators, communities, even legislators should be able to agree on weak areas for improvement. They should be able to agree on appropriate outcome measures to monitor improvement across time. They may debate but generally will agree on an expected level of performance, long-term. Tracking the chosen outcomes from the first year of measurement across time should not be an arduous task. All concerned would be playing by the same rules and looking for the same results.

The concerns regarding interpretation of the information on the school report cards are embodied in the following questions.

Do data lend themselves to clear interpretations of success/failure? How "bad" is a grade of 82 compared to a 96? How "bad" is a change of one percentage point from one year to the next?

Is good/bad clearly discernible? Would a jury of "twelve peers" agree on good versus bad inferences? Are good and bad measured on an equal-interval scale?

Are confounding factors clearly identified? Are confounding factors accounted for ahead of assessment and evaluation? Are all necessary assumptions met for the statistical method chosen? Are process

variables accounted for in the analytical design? Are fundamental subpopulation differences taken into account in grouping strategies?

Are presumed cause-and-effect relationships valid? Correlation does not imply causality. Excellent scores may not imply great teaching and administration. Similarly, poor or average scores may also not imply less than excellent teaching and administration. Process analysis provides this information.

How will practitioners use the results to improve instructional practices? Can a teacher or principal look at the report card and decide what specifically must be done to change any of the outcome measures in the next year? Are any specific guidelines available from TEA to assist teachers and principals in translating from the report card to action plans?

The Texas AEIS: Outcomes, Intended and Unintended

Once they hit the newstand, the Texas school report cards will undoubtedly generate heated debate for weeks. Perhaps the most sobering aspect of this particular accountability experiment is the high stakes attached to success and failure within the community. Will the report cards help local communities determine the status of their public schools, or will they create a smoke screen for displaced aggression among a troubled populace?

Parents' right to choose their child's school, regardless of attendance area, is inevitable, according to the opinions of many. It is not clear whether choice will impact public education in the way free enterprise affects business and industry. Taxpayers certainly have a right to expect, even demand, quality for their contribution to the public schools. Are parents contented to support a local school financially that gains a reputation for being a substandard campus from the report cards? Will they select an alternative school based on the same criteria — the report card statistics? That we must invest our resources in the most effective solution is a given. That the most effective solution has been identified is debatable, as is the method for identification.

Districts stand to lose considerable funding if vouchers become law. The likelihood of passing bond initiatives in many districts may become less than zero after the AEIS statistics are released, whether the data are accurate or not. Rollbacks are a possibility for other districts. It remains to be seen how many districts can provide administrative support services to schools on budgets that are drastically impacted by these factors. If districts cannot provide some of these services to the campuses, then who will?

What is the likely impact of these events upon the morale and efficacy of faculty and staffs at the schools rated low on the report cards? Many teachers and administrators already suffer from burnout, as they labor in a profession that fails to command the respect of the community and commensurate financial compensation.

Educators have been leaving the profession in growing proportions for the last two decades, moving into business and industry where salaries and prestige provide powerful incentives. The most highly skilled educators are the ones who have no trouble beginning alternative careers. Unfortunately, some of the less skilled educators remain in the schools. Trimming dead wood is always beneficial. Wood trimming is a U-shaped function of returns. Plus, one must take care to trim only the dead wood and not whack out

portions of vibrant growth in the process. How do we as a nation expect the most highly skilled, brightest graduates to assume educational careers, when being a professional educator day in and day out is so unpleasant?

What must students think about educators, schools, education as a profession, and even the enterprise of learning? Will they be able to maintain a serious attitude towards it? Are most students motivated to attend every day because of the joy of learning and the fascinating character of their teachers and coursework? As they observe the debates regarding the "F" schools and the "A" schools, what will be their reactions?

Agenda for Improvement

Table 2 summarizes the issues targeted for reconsideration in the AEIS and school report cards. Improvements fall into four primary categories: basic design components, grouping issues, accountability measurement, and training.

In terms of basic design, the indicators for socioeconomic status and mobility deserve further investigation and refinement. TEA can improve the system and its acceptance partially by designing a procedure for statewide use in tracking, collecting, and reporting mobility. Related to statewide procedures, there is widespread skepticism about the absolute accuracy of many of the variables contained in TEA's database.[2] Quality-control procedures need to be developed and published to ensure the integrity of the database and the validity of the conclusions drawn from the analysis of the variables.

A pilot of the entire system is needed before publication of the results. All aspects of the AEIS and report cards should be evaluated in a field trial: grouping and accountability factors and methodologies, presentation format, locally referenced objectives and progress monitoring, and training in interpretation of the results.

Schools do not learn; *individual students* learn. Student-referenced progress towards goals would serve as the basis for campus evaluation. The pilot could evaluate the use of a campus-referenced model for monitoring and accountability, rather than a statistical group of 100 "like" schools. Elements of the pilot which are successful would be *gradually* introduced into the system annually, instead of allowing the system to evolve continuously through trial and error. This enables model stability for the sake of consistent comparisons across time.

Grouping strategies are problematic. The variables and statistical procedures used currently with the report cards are overaggregated and insensitive to salient differences among campuses. Two or three dozen grouping variables are unnecessary, but greater sensitivity and less redundancy among them is desired. If some of them were manipulable by local communities and school staffs, perhaps real experimentation would occur based upon better monitoring tactics. Statewide comparisons are useless, if the campuses compared are dissimilar in terms of fundamental processes and deterministic local factors.

---

[2] Public Education Information Management System (PEIMS).

TABLE 2

IMPROVEMENTS NEEDED IN AEIS AND TEXAS SCHOOL REPORT CARDS

DESIGN AND SYSTEM IMPROVEMENTS

Better Procedures
- Tracking and collecting mobility data
- Campus-referenced evaluation (not normative)
- Built upon individual[student]-referenced assessment

Better Definition
- Mobility index
- Socioeconomic indicator(s)

Increased Quality Control
- Data accuracy
- Pilot of system, followed by thoughtful revision
- Introduction of successful pilot components gradually rather than annual changes to the system
- Model stability across time

GROUPING FOR COMPARISON

Indicators
- More diverse - spread out population
- Some within experimental control

Methodologies (statistics)
- Confirmatory (more than one method)
- Truly create "like" clusters of campuses, according to impacted population
- Involve impacted populations in design and formative evaluation
  - *Students*
  - *Parents*
  - *Business community*

ACCOUNTABILITY MEASUREMENT

Outcomes
- Need process indicators and strategies for measuring them
- Improve product indicators
  - Related to "real world" success
  - Related to long-term success after high school graduation
  - More local input
    - *Students*
    - *Parents*
    - *Business community*
    - *Site-based leadership team*
    - *District*

Outcome Methodologies (statistics)
- Confirmatory (more than one method)
- Dissaggregated results by appropriate subgroups

Comparisons
- Longitudinal
- Within campus
- If across campuses, need better groups/clusters

TRAINING

For Interpretation by Nontechnical Groups

For Interpretation by Educators

Guidelines for "Significance" of Changes
- Year to year within a campus
- Year to year campus to group
- At all levels of every cluster
- Monitor for drift

Guidelines for Campus Staff for Prescriptive Use of Results

The best way to facilitate sensitive monitoring is to involve people most familiar with the campus — parents, students, faculty, staff, and local community members. This premise has been proven repeatedly in the corporate environment with total-quality management techniques.The accountability aspect of the system is too limited and biased toward available standardized test data. Process factors have been completely overlooked. Product measures are hardly related to real-world success after graduation — employability, success with further education, on-the-job evaluations, functionality in interpersonal relationships (family and otherwise), and the rights and responsibilities of citizenship, to name a few.

Statistical procedures for prediction of success in reaching goals bear review. Perhaps we in school evaluation have become habituated to the hackneyed, uninventive, and overused model of the annual comparison of group averages. We have used this method for so many years in so many different circumstances to "evaluate" achievement. Is there no hope for appropriate alternatives? Is regression the absolute height of ingenuity in our profession? Can no other descriptive statistical technique offer a fresh point of reference?

The report card information must be longitudinal if it is to be used from year to year in a meaningful way to enact operational changes. This implies that the underlying model is consistent and stable. A campus comparing itself to its own performance across time stands a much better chance of pinpointing areas for improvement, than an overaggregated comparison with a number of not-necessarily-homogeneous campuses statewide.

The issue of training the users (and those *used by*) the AEIS and the school report cards has been woefully neglected. A seventeen-page publication explaining variable definitions, data sources, and other qualifying instructions for the report card is wholly insufficient as a training tool to a nontechnical public. In addition, there are no guidelines for school staffs suggesting how to use the information for school improvement. Some sort of analysis of the significance of varying increments of change is lacking, both between campuses and group averages and within a campus across time. Simply handing the report card to a school is not going to assist them in any way to know what it means or how to proceed. No planning has occurred that addresses the realistic amount of time required for both the nontechnical public and the professionals in education to be able to interpret the AEIS information, even if training materials were available. Widespread *accurate* understanding seems to be improbable, under the circumstances.

Can Anyone Out There Tell A Trunk From A Tail?

One major complaint about current assessment systems is the relationship of measured outcomes to instruction. Norm-referenced tests by definition spread test takers out in terms of performance, and items out in terms of difficulty. These tests were never intended to pinpoint an individual student's ability, or sample *with precision* a student's comprehension of a unidimensional, clearly delimited construct of learning.

Many professionals in education today are jumping on the bandwagon for alternative assessment as the replacement for traditional norm-referenced or criterion-referenced testing programs. A thoughtful treatment of the pro's and con's of

snubbing traditional methods and switching to more "authentic" (i.e., "old-fashioned") methods of assessment appeared in the December 1991 edition of *Phi Delta Kappan*.[3] In that article, Maeroff reminds us that speed and low cost have always been the main selling points of norm-referenced testing. In fact, the entire tradition of norm-referenced tests answered the cry for efficient, quick, cheap, and manageable assessment of *masses* of test-takers.

Norm-referenced testing is a paradigm designed to yield generic descriptions of individual abilities and aptitudes easily and quickly, and generic profiles for large groups of test-takers based upon clearly predefined, standardized factors of the testing situation. The tradition started with the U.S. Armed Services requirements in World War II for immediate assessment of incoming recruits. It gathered momentum in public education as the baby-boom generation hit the American schools like a tidal wave.

In the beginning of public education in America, teachers understood that individuals were to be *taught and evaluated as individuals*. It never occurred to them that individual students could (or *should*) be lumped together as a group of "learners" and represented by one number in terms of their concerted efforts. In fact, such a proposal would likely have struck pioneer teachers as immanently ludicrous. How can we expect to identify learners' strengths and weaknesses and bring each one's level of comprehension to a high, common standard *without dealing with each student on a one-to-one basis*? To be sure, that was time-consuming. Teachers made up individual tests for each student, unique homework assignments, and graded them all as a regular matter of course. When we consider the amount of knowledge and information to be taught before 1920 compared to today, perhaps the task is not as terrifying as it sounds in today's context.

Maeroff points out that alternative assessment is generically time-consuming and labor-intensive. He claims that it is sometimes "imprecise" in terms of inter-rater reliability for ratings awarded. However, there are elements of alternative assessment which pioneer teachers would likely applaud. One is the fact that this type of assessment is *embedded* in instruction. Maeroff comments that when assessment is "sheared off from instruction," as it is with norm-referenced tests, results are generally unrelated to learning and classroom performance. He argues for a "semester-length video" as opposed to a "Polaroid snapshot" of learner accomplishments.

Are teachers affected by the snapshot approach to evaluation? Are there unintended outcomes of this type of assessment which pose obstacles to educational reform? Gerald Bracey summarized several studies in the November 1991 issue of *Phi Delta Kappan* pertaining to the results of publishing large-scale averages on norm-referenced testing.[4] His comments describe the current state of mind in Texas. Bracey quotes a study by Halladnya and Haas, which defines *test-score pollution*, describing the situation in which there is a breakdown in validity that underminds inferences drawn from testing. Test-score

[3] "Assessing Alternative Assessment," Gene I. Maeroff, *Phi Delta Kappan*, December 1991, vol. 73, no. 4, pp. 272-281.

[4] "Research," Gerald W. Bracey, *Phi Delta Kappan*, November 1991, vol. 73, no. 3, p. 255.

pollution is especially prevalent in high-stakes situations, as is the case with the AEIS and Texas school report cards. The researchers claim that educational reform is unlikely as long as test results *misrepresent* American public education in this way. Bracey summarizes it in his "paradox": "Test scores mean something only when you don't pay attention to them."

Bracey further quotes from a study by Smith,[5] in which teachers' responses to publishing students' test scores were surveyed. In summary, the following outcomes were not surprisingly observed:

- Due to feelings of shame, guilt, embarrassment, anger, stress, and anxiety, teachers are likely to "do whatever is necessary" to avoid these feelings under similar circumstances in the future.
- Teachers believe the scores are invalid and yet they must raise them on future assessments. This creates a painful state of dissonance.
- Teachers also believe that young children are emotionally impacted by their experiences with and scores on tests. This creates anxiety and guilt for the teachers.
- Tests reduce time for instruction.
- Tests narrow the curriculum through focus on coverage of the areas included in the tests.
- Multiple-choice testing leads to an attenuated set of teaching techniques. This leads to a much-reduced pedagogical-skills set among teachers, and less acquaintance with higher-order thinking skills in regular classroom instruction.

The result is the common observation among today's classroom. One result is burned-out teachers. Another is a test-driven curriculum resulting from high-stakes assessment predicated on a reduced set of both objectives and requisite thinking skills.

President George Bush announced his administration's goals for America 2000 in April of 1991. Reactions were sought of a diverse group of leaders, evaluators, and close observers of previous federal afforts to improve schooling and education. The resulting collections of professional responses to America 2000 were published by the William T. Grant Foundation under the title of *Voices From the Field.*[6] Thirty experts provided their reactions to the philosophy behind the America 2000 goals and the likelihood of such a program coming to fruition, given the current status of education in America.

In *Voices From the Field,* Linda Darling-Hammond reminds us that norm-referenced tests were initially created to track and sort large numbers of students efficiently. They were never intended as tools for instructional enhancement. She continues that higher-order thinking skills have been steadily decreasing in classroom instruction, while basic-skills instruction has increased since 1970. She claims that "...assessment cannot be a constructive lever for reform unless we invest in more educationally useful and valid measures of student learning."[7] Hammond gives voice to an observation many of us have shared

[5] *Ibid.*, p. 256.

[6] *Voices From the Field*, William T. Grant Foundation Commission on Work, Family, and Citizenship, Washington, D.C., 1991.

[7] "Measuring Schools is not the Same as Improving Them," Linda Darling-Hammond, *Voices From the Field,* pp. 15-16.

in the last two decades, namely that "...assessments of any kind will not be sufficient to stimulate all the changes and improvements America needs and wants from her schools." If this is true, education will not be the savior for societal reform.

Gary Orfield and Jeannie Oakes enumerate other negative consequences of test-driven curriculum and reform movements. Orfield[8] lists the arguments against test-driven curriculum as costs in time and money of *one more* layer of tests, inevitable cultural bias, and prevalent misleading use of scores, by technical and nontechnical audiences alike.

Oakes[9] focuses our attention on the impact of test-driven reforms on students, especially those of low socioeconomic and minority subpopulations. She argues that Latino and African-American students are particularly at jeopardy in this gambit. She claims that the uneven distribution of resources and opportunities is predictable if not guaranteed to those disadvantaged by low socioeconomic conditions and minority status. In addition, test results are very often used to *close off* rather than *open up* future opportunities for low scorers in these categories. She seems to interpret the current movement towards the system of nationally assessing performance in "break-the-mold schools" as a formidable obstacle to the wholesale guarantee to a "break-the-mold education" for a growing proportion of our youth.

Two writers surveyed in *Voices From the Field* provide succinct insight to the reasons for the inevitable failure of accountability designs such as the school report cards.

> "The proposed 'Report Cards' reflect ... narrowness of thought. Publicized results could stimulate public discussion about how to improve educational outcomes, but, as proposed, they are more likely to incite unfair comparisons. What can be tested easily and efficiently is a small part of what the fully educated person is and falls short of what the technological world requires to keep America on top."[10]

> "Comparisons of schools, districts, and states will demonstrate what is already known: that some segments of the population perform deplorably in the education process ...Schools with large numbers and concentrations of minority, disadvantaged, and other atypical populations will do worse than schools enrolling White, middle-class, English-language background students. Central city and rural districts will perform more poorly than suburban districts. States with homogeneous White populations will outperform states with heterogeneous populations."[11]

Fairness or appropriateness of *comparisons* is critical to the success or failure of not only assessment but also reform efforts. This is a major criticism cited earlier of the AEIS and report cards. A report authored by C. C. Carson, R. M. Huelskamp, and T. D. Woodall[12] cites aggregated

8 "Choice, Testing, and the Re-election of a President," Gary Orfield, *ibid.*, pp. 7-8.

9 *"The Many-Sided Dilemmas of Testing,"* Jeannie Oakes, *ibid.*, pp. 17-18.

10 "Steaming Backward to 2000," William W. Wayson, *ibid.*, pp. 48-49.

11 "Widening, Not Narrowing, the Gap," Jose A. Cardenas, *ibid.*, pp. 28-29.

12 "Perspectives on Education in America: Annotated Briefing, Third Draft, May 10, 1991,"C.C.

(and *skillfully disaggregated*) data that support this notion. They summarize a rehash of data from various sources (National Center for Educational Statistics, National Assessment of Educational Progress, Digest of Education Statistics, U. S. Census, Educational Testing Services, The College Board, Bureau of Labor Statistics, Hudson Institute, and other national data clearinghouses). Their analysis, however, is distinguished by *disaggregation* of the information based upon significant demographic factors. Some of the details from their briefing will be cited later. For now, the focus is on their summative commentary regarding *quality of data* used in evaluating the educational system in America.

They conclude that the data used for determination of accountability of public education in America are of *very poor quality*. They conjecture that this is "due to the [typically] insufficient funding of data collection, synthesis, and analysis at all levels of education." This is true in Texas, where high-stakes decisions are made almost daily based on numbers drawn from the TEA database, yet no budgetary allowance or staffing guidelines have ever emerged from the State to ensure the accuracy and validity of the data.

Carson et al. surmise that "evaluators" often "fall back on whatever data are available, regardless of quality." Many veterans in evaluation may reluctantly verify this hypothesis in many cases. Carson et al. continue that the "use of overly simplified or highly aggregated data will result in poorly focused actions with disappointing outcomes." This prediction may be coming true in Texas.

The authors further contend that the "misuse of simplistic measures with dubious value leads to reform obstacles." When the media grab hold of school report cards and begin to compare campus to campus within a community, public perceptions are affected. To the extent that the media succeed in disenfranchising the taxpayers regarding their public schools, the odds of real reform occurring *within a community* to bolster education decline drastically. Alarm calls focused on falling SAT scores and unfavorable comparisons of aggregated American averages with those of other cultures foment less than a supportive attitude among readers. Carson et al. comment that professional educators are hesitant to use current educational data to support their calls for reform. This is due to flaws in available data and the trend of complex data being reduced to absurdly meaningless, overaggregated,summarystatistics.

Table 3 summarizes selected dropout statistics published by Carson et al. Dropout rates are decreasing for all ethnicities except Hispanics. The reason for the continuing high dropout rate for Hispanics is that 50% of those of school age in that ethnic group are first-generation immigrants. Most of these never attended an American school, yet they are counted in the raw numbers of dropouts by ethnic group. Carson et al. say that among Hispanics born in America, the dropout rate is about the same as that for Blacks.

Inner cities have the worst dropout rates, suburbs have the best. Black students in suburbs show the smallest dropout rate of any ethnic group and any community type.

Carson, R. M. Huelskamp, and T. D. Woodall, Systems Analysis Department, Sandia National Laboratories, Albuquerque, New Mexico.

TABLE 3 DROPOUT RATE STATISTICS	
BY ETHNICITY	Has decreased for Blacks by nearly 50% in the last twenty years (27% to 15%) to currently within 2% of that for Whites. Has not declined for Hispanics but is heavily impacted by: • High immigration rates for Hispanics. • Fifty percent of school-age Hispanics are first-generation immigrants, most of whom never attended American schools.
BY COMMUNITY TYPE AND ETHNICITY	Declining for all types of communities and races combined with exception of Hispanics. Highest for inner-cities regardless of race. Highest for Hispanics regardless of community type. Lowest for suburban Blacks..
BY DEMOGRAPHIC VARIABLES THOUGHT TO BE AT-RISK INDICATORS	At-risk traits not typically exhibited, such as minority membership, limited-English proficiency, retained at grade level at least once, raised in single-parent home, married, or a teenage parent. Nearly twice as many children since 1960 raised in a home where both parents work. More than twice as many children since 1960 raised by a single parent. Proportionally, nearly twice as many children raised in single-parent homes drop out as those raised in two-parent homes (1.55 vs. .85). More than twice as many divorces today compared to 1960. Nearly twice as many (40%) children raised by one parent today compared to 1960.
BY REGIONAL DATA	No correlation whatsoever between dropout rate and any other parameter collected at district level. Patterns reflect state-by-state differences and needs. Students leave school early for a variety of complex reasons.

It turns out that the majority of dropouts do not represent the at-risk stereotype. One concerning at-risk indicator, however, is that of single-parent homes. Children from single-parent homes are not overrepresented in the statistics for dropouts. However, they do drop out at a rate nearly twice as high of that of children from two-parent homes. Given the rising divorce rate in America, this may prove to be a serious jeopardizing factor among already at-risk populations in the future. Any solutions to decrease dropout rates must address separately the disproportionate rates among ethnicities and the predisposing factors towards dropping out, regardless of ethnicity and at-risk status.

Carson et al. review the role of immigration in the demographic and academic performance changes within the last 30 years. Table 4 summarizes their data describing trends related to legal and illegal immigration. Predominantly Europeans immigrated to America after World War II. During the 1980s the majority switched to Latino and Asian, especially when illegal immigrants and refugees are counted. Recent legislation will decrease the 85% Latino and Asian influx to somewhere between 60% and 70%. However, this does not account for illegal immigrants and refugees, who most likely will

represent these two ethnic categories. With over 150 languages nationally, how will the educational system cope with this terrific heterogeneity *and* guarantee a world-class education to all?

TABLE 4	
IMMIGRATION TRENDS	
LEGAL	Highest rates since 1920. Accounts for nearly 30% population growth in America in 1980s. Accounts for demographic shifts in schools and work force. Limit increasing by 17% (to (700,000 per year) due to legislation. Majority in 1990s anticipated African and Eastern European.
ILLEGAL	Estimated 200,000 per year in 1980s. Estimated 100,000 per year in 1990s. Estimated additional 130,000 refugees in 1990s.
BY ETHNICITY	Mostly European after World War II. Mostly Latino and Asian now. Legislation decreasing proportional rates of Latino and Asian, but most illegal immigrants come from these two groups. Estimated up to 5 million immigrant children in American schools in 1990s. More than 150 languages nationwide.

TABLE 5
FACTS ABOUT THE WORK FORCE IN YEAR 2000
Over 70% of year 2000 work force currently employed.
Additional 7% of year 2000 work force immigrating in 1990s.
Remaining 22% of year 2000 work force coming from American schools.
Current educational reforms impact only 20% of year 2000 work force, most of whom are now in later stages of education and will be entry-level, younger workers then.
Most of impact from need for higher skills is on currently employed workers.
Traits wanted most of all by labor: drug-free, respectful, punctual, honest, basic reading ability for directions and forms — generally behavioral not academic.

Carson et al. also describe trends related to SAT averages and test-taker characteristics. According to self-report data collected on the SAT, there are less students ranked in the top 20% of their class taking the SAT now than 15 years ago. At the same time, there are more students ranked in the bottom 60% of their class taking the SAT *than ever before*. This is good news in terms of increasing access to postsecondary education for all students, not just the top of the class. However, it does raise serious concerns about the overaggregated use of SAT and ACT averages as indicators for campus accountability.

In a previous section of this report, a plea was made for more outcomes that represent the real-world better than test scores. Carson et al. show with their data that a little more than one in five employable workers in the year 2000 are now attending public school in kindergarten to grade twelve. This leaves the vast majority of the work force, which we hope will excel in world-class competition, either already employed (i.e., out of school) or waiting to immigrate into the country. Thus, current educational reforms will impact most of all only those public education students currently within the later stages of their education.

If business and industry are asked what skills they want in their new employees, they generally cite skills in the behavioral domain, not academic prowess. Table 5 summarizes a few facts in this area. They want respectful, punctual, honest, and drug-free employees. Reading is a needed skill, for the ability to understand instructions and read signs and other simple material.

Maybe business and industry are not as concerned after all with the average SAT and ACT scores of locally graduating seniors. This does not mean that college-graduated employees are unnecessary. It does suggest that more outcomes are needed which address more than just eligibility for college acceptance.

In the December 1991 issue of *Educational Researcher*, Robert Hess and Hiroshi Azuma[13] report cultural differences which may help to account for some of the variance among American and Japanese students in school performance. Hess and Azuma point out the school is not "user-friendly," and that much socialization is needed to increase a child's probability of success in this environment.

Japanese mothers in their study consistently attested to the importance of compliance, cooperation, diligence, patience, and obedience in their upbringing of children. American mothers, on the other hand, valued most of all independence, tolerance of others' opinions, and initiative. No wonder that Japanese children stick to the task assigned by their teachers and persevere.

Our cultures prize different behavioral styles from youngsters as they grow and learn. Hess and Azuma point out that creating a microculture within a school to counter the effects of society at large while encouraging alternative values is a difficult assignment. More than just administrative and curricular changes would be needed to accomplish such an environment. These types of factors illustrate those process indicators which are absent from the current report card design.

---

[13] "Cultural Support for Schooling: Contrasts Between Japanese and the U. S.," *Educational Researcher*, Robert Hess and Hiroshi Azuma, December 1991, vol. 20, no. 9, pp. 2-8.

Taking Off the Blindfold

We must design new methods for measuring accountability other than group averages and wall charts. If we truly believe that every child can learn, then why look at group averages? The Texas school report card is an attempt to determine accountability. However, it places ultimate priority on simplicity of definition and statistical methodology. This plus the selection of grouping factors and performance indicators render it too gross a measure for the task. All evaluation must start at the student level of information. This is our current starting point, but the aggregated measures derived from this data confound too many critical factors.

Don Glines and Kathleen Long[14] recently enumerated thirty-two transitional steps which they feel will lead to the American school of the future. Strikingly, many of their recommendations were echoed in the over 800 proposals recently submitted to the New American Schools Development Corporation in February of 1992. These proposals compete for private funding to design the "break-the-mold school" of the twenty-first century, in answer to President Bush's goals expressed in America 2000.

Glines and Long summarize changes which must occur in assessment in order for American schools to succeed in the next century. They see committees meeting weekly comprised of parents, students, and staff members to determine goals, interests, and needs of students. The committee would likely function similarly to a medical staff meeting.

Assessment should be individual-referenced, not norm-referenced. There should be no report cards (for individual students), no grades. An arsenal of alternative assessment strategies would be used (for example, portfolios and peer assessments) to monitor individual progress towards individual goals. Program evaluation "... would emerge as part of the planning process." Certainly, these changes are supportive of Maeroff's[15] conclusion that, since classroom teachers are the assessors of their students, they should be formidably armed to produce the best information.

The National Center for Research on Evaluation, Standards, and Student Testing (CRESST) recently published its task force's recommendations on national assessment.[16] They emphasize that "break-the-mold" assessments are needed to measure the national educational goals. The eventual system should fulfill their vision of a *positive role of assessment in educational reform.* Much assistance will be needed, both monetary and technical resources, from all levels of administration in order to achieve this goal. The following quote summarizes their recommendations for technical assistance, development, and research:

> "Finally, a research program should be funded to solve problems in design, analysis, and interpretation of new assessments. Areas requiring research should include, among others, new approaches to comparability, generalizability and fairness of

14 "Transitioning Towards Educational Futures.", *Phi Delta Kappan*, Don Glines and Kathleen Long, March 1992, vol. 73, no. 7, pp. 557-560.

15 *Op. cit.*, Maeroff, p. 276.

16 "Task Force Recommends New Assessment System," *The CRESST Line*, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Winter 1992, pp. 5-6.

assessments, adaptation of assessments to special needs of students, and issues of aggregation and reporting." (p. 6)

What we need is nothing short of a paradigm shift in individual assessment and educational evaluation. We need individually relevant goals and goal-specific measurement. Outcomes for individuals can be tallied to provide a holistic measure of campus performance. Roland S. Barth[17] defines a paradigm shift as licking the envelope and not the stamp, if the taste of the stamp is really unbearable. This provides an appropriate metaphor for turning our attention from the class and school to the student. After all, it is the envelope and its destination which is of utmost concern.

Barth poses tough and insightful questions about the truths that faculty and school staffs must face in order to assess student and school needs, which will lead to goals, which will chart their communal course for the future.

He reminds us that the only person that enjoys change is a wet baby. This metaphor is not wholly inappropriate, considering the cries and screams coming from political quarters about the status of American education. Massive staff development is mandatory among federal, state, university, and district ranks in order to pave the way to success for school-initiated reforms. We must recognize that those who work individually with students — teachers, parents, peers, and school staffs — deserve to be inspired by their own mutual vision rather than the prescriptions of others. Barth reminds us of Marcel Proust's observation that the "...real art of discovery lies not in finding new lands, but in seeing with new eyes."

In order to see with new eyes, we must remove our blindfolds. Campus averages, unless they are disaggregated to the point that they no longer represent whole campuses, must be relegated to their proper place in evaluation.

We are attempting to measure *learning*. Districts do not learn. Schools do not learn. We should focus on individual-referenced needs, goals, and measurement of progress. We have no other choice, operationally, in a system where education is tailored to meet the needs of each learner. How can we justify measuring a system's success in this endeavor in any other way except aggregating success stories? When group means are calculated, successes and failures *and near-misses* are thrown in together. Just as when a washer contains blue jeans along with white gym shorts and socks, the result does not fairly represent any of the individual elements, nor the cleaning equipment or process.

Barth quotes the original German definition of *professor* as "one who thinks otherwise." It is time for us to think otherwise, to expand our vision outside of the nine-dot box and focus on the goals of education in America *as it is evolving in the next twenty to fifty years*. If there is indeed strength in diversity, then we must have all we need to exert our claim to the number-one position in world-class comparisons.

Business and industry have two particular paradigms to offer us which may show the way to our future success. One is *total-quality management*, which to a large extent is just emerging in education in site-based leadership practices. More frequent practical applications of total-quality management to edu-

17 "Restructuring Schools: Some Questions for Teachers and Principals," Roland S. Barth, *Phi Delta Kappan*, October 1991, vol. 73, no. 2, pp. 123-128.

cational settings *and subsequent reports* will serve to enlighten us all regarding our options and potentials.

Another paradigm which has revolutionized business is *benchmarking*. In order to *be the best*, businesses find out who *is the best*. They study the top-of-the-line producer and then mimic them functionally. Benchmarking is specific to *functions* performed in the business, such as customer service, training, and inventory of parts and supplies. Comparisons are not restricted to within the industry.

If a social studies department in Abilene, Texas, determines that the best mapmakers are in the U. S. Navy, they could research as much as possible through information networks how the Navy "does it." This introduces one of the side benefits of benchmarking. It is an unobtrusive device for (a) networking and (b) evoking business and community support of schools. At some point, the department chair and teachers would communicate with the appropriate division within the U. S. Navy and ask for further information. An on-site research team from the school might even go and observe the servicemen at their work, interviewing them and getting more specific information they can take back to Abilene and instill in the social studies program.

Why can't teachers and students design instructional goals and programs in the same way? Again, staff development (retraining) proves itself to be a major requirement for change. Teachers need to know the principles and strategies of benchmarking. This particular activity applies directly to collaborative and cooperative learning, as well as to action-oriented, real-life instructional methods. Students would learn the technique of benchmarking from experiencing it in school, through teachers, peers, and mentors. Parents could learn and apply the techniques through parent education classes at school, or through involvement with their childrens' learning activities. Within two generations, we could disseminate the principles of benchmarking throughout our society. As a matter of fact, benchmarking is partly a formalized version of apprenticeship and self-paced learning. Think of the result of producing a generation and a half of benchmarking citizens by 2030. That really is a vision.

Dr. Michael Hammer of Boston is currently presenting *re-engineering* seminars throughout the country. This is a new technique which business and industry is devouring as a means to world-class competitive status. In re-engineering, the workers most familiar with the basic functions of the industry form a redesign team. In the process, the entire system reflected by the business in its involvement in the industry is redesigned. The criteria of high customer satisfaction, high profits, high worker satisfaction (extrinsic and intrinsic), and high cost-containment and efficiency are met. The application of this method to educational assessment, at the individual and large-group levels, could result in major benefits to all parties involved — students, parents, schools, communities, districts, states, and federal agencies. Efforts to investigate the viability of re-engineering in the educational setting are currently underway.

Now That We *Know* It Has Tusks,...

The Texas school report cards do not provide the quality or type of accountability information for the tough decisions facing us today in education. Comparability among campuses is suspect, based on both variables and methods used. Out-

come measures are not comprehensive, nor are they sensitive to the processes operating in our schools. Outcomes are also biased too heavily towards overaggregated test scores. The tests included are hardly representative of the curriculum taught in Texas schools, and certainly not of real-life skills desired by students and community members.

The evaluation and assessment system in public education must become more individual-referenced. It must rely on sensitive assessment to detect individual needs in a variety of areas, all of which are not simply academic. We need to define the "learner" as one who needs to know, and then proceed with the identification of what that encompasses for each learner.

Needs may shift as our nation and the world changes. Our system must be flexible enough to accommodate evolution *and revolution*, and not incarcerate itself in outdated ideas of personal excellence.

Accountability for programs, schools, and teachers must reflect the actual functions and goals in operation for the benefit *of each learner*. How else to judge success except to look at the avowed product and its quality? Otherwise, we risk washing out the results (*good* or *bad*) for individuals in the same way we historically diluted the results for minority groups. Has the individual learner finally become a minority in our educational system?

In order to get away from the "basic-skills mentality," we must acknowledge the critical role of higher-order thinking skills *and accommodate our instructional and measurement systems* to focus on these with a priority on accuracy. Results must be indexed to significant factors additional to those currently sought, such as *accurate* mobility indicators, *realistic and valid* socioeconomic measures, length of enrollment in current program, and parental influences, to name a few.

Nationally, experts in educational research and evaluation seem to be converging towards a consensus. Goals may be set nationally. The federal Department of Education, in conjunction with appropriate advisory groups, may define "world-class standards" for all schools and students. Local schools and communities should then take the responsibility for casting national goals in terms of individual and community needs, interests, and activities.

It is up to the federal and state agencies to *monitor the progress* of each school towards the national goals. State and federal governments implicitly establish local priorities when they mandate assessment and accountability systems. Rather, they should specify where we are all headed in terms of curriculum coverage and quality. Leave it to the local communities and their schools to reach the goals by their own best judgement and efforts.

In other words, just tell the local schools we are looking for an elephant, and that it is supposed to be a big one. They need to know how to tell how "big" is "big enough." They may need other details so that they can get the *vision* of the desired outcome. Once they have that vision, however, schools and communities are more than capable enough, together, to reach the goals.

This may mean opening a can of worms. President Bush's America 2000 goals have sparked an interest in communities all around the country. Mentioned previously are the 800+ proposals submitted to design the "break-the-mold school" of the twenty-first century. What may not be well

publicized *yet* is the all-consuming involvement demonstrated by business, industry, parents, and communities at large in discussions reacting to the request-for-proposal from the New American Schools Develoment Corporation. Whether or not proposals are funded *or were even submitted*, the call to arms has made itself manifest to the American public.

It is not, really, that we Americans are competitive by nature. Rather, it is that we hate to lose. We are not good at losing. We sulk and pout and tell how "it would have been different if..." over and over, to any who will listen.

The truth is, we are splendid winners. Charitable, forgiving, helpful almost to a point of dysfunctionality. We share any wealth we glean with any who are destitute. We educate the masses from other countries. Yes, they pay tuition to our colleges and universities. However, we do not *have to* admit them to programs. In fact, there is no reason not to admit eligible students. We as a people believe in the inalienable right of an individual to education, information, and the access to personal growth. So much the better if she or he turns around and remits to society something of the legacy earned.

We are a nation committed to the betterment of *all people*. This has been the case since our incipience. Admittedly, the path to betterment for many has taken intriguing if not downright spurious detours. However, a nation that believes, through education and training, that everyone can learn, everyone has something valuable to offer, and that the best solutions meet the needs of those most impacted ....

Well, whether or not it is possible, I am convinced *for myself* that it is worth investigating.

And, isn't that where the choice to change starts?